研　究　報　告　書

「解析過程と応用を重視した再利用が容易な言語処理の実現」

研究タイプ：通常型
研究期間：　平成 23 年 10 月～平成 27 年 3 月
研　究　者：　狩野　芳伸

1．　研究のねらい

本研究のねらいは、解析過程と応用を重視した自然言語処理技術を開発することである。

近年、産業界でも自然言語処理技術の需要が高まっている。しかし、自然言語処理技術の利用には高度な専門知識を持ったプログラミング作業が必要であった。本研究では、専門外でもより多くのユーザが専門的な自然言語処理技術を利用できるようになることを目指し、利用に際して必要となる様々な作業の自動化を支援する実行プラットフォームを構築する。プラットフォームは、データやツールの組合せ、インストール、実行、評価、分析（統計・視覚化）といった機能を、国際標準 UIMA に準拠しつつ統一的な実行プラットフォームとして極力汎用になるよう実装する。また、プラットフォーム自体を任意のマシンで実行可能とし、プラットフォームを始めて使う時点からの徹底した自動化プラットフォームを目指す。

プラットフォームの開発と同時に、新規に、あるいは既存のツールやデータを再利用の容易な形に再構成したうえで、UIMA 準拠かつプラットフォーム互換のツールキットとして統合する。ツールキットに関しても、極力任意のマシンで実行できる可搬な実装とするとともに、プラットフォームにおける自動組み合わせ機能に対応できるよう、適切な入出力定義をする。入出力のデータ型定義は、将来の発展を考慮して汎用かつ体系的な階層をなすようにする。

並行して、より自然な動作ができる言語処理ツールの研究を行う。既存の言語処理ツールの多くは、ツール自体がどれほど人間の作成した正解と一致するかという性能評価を向上させるべくチューニングされている。その一方で、ツールの解析過程は人間のそれとは異なったものになっている。そこで、人間の言語処理における解析過程の知見を取り込んだモデルの構築と、プラットフォームでの応用評価を目指す。

総合すると、質問応答から音声認識まで、さまざまな互換ツールを標準ツールキットとしてプラットフォームとともに配備し、実用的アプリケーションでの応用を試みる。ツールの配備はオープンアーキテクチャとし、第三者も容易に配備の追加ができるようにする。全体として、世界に類を見ない統合全自動言語処理システムを構築するのが狙いである。

2．　研究成果

（1）概要

プラットフォームの構築は、おおむね狙い通りのものが完成した。すなわち、データやツールの組合せによる自動ワークフロー生成、自動インストール、自動実行、メトリクスのカスタマイズ可能な比較評価統計、汎用視覚化といった機能を、国際標準 UIMA に準拠しつつ統一的な実行プラットフォーム”Kachako”として構築した。また当初計画にはなかったが、共同研究を通じ、テキストとアノテーションの同時検索システムの実装を行い、統合システムの出力を汎用的なクエリで検索可能とした。

ツールキットの構築においては、質問応答・音声認識・大学入試自動解答・医薬品文書解析など、当初予定にはなかったさまざまなツールへの対応も含め、多くの共同研究を通じて達成することができた。また、大学入試の自動解答では最高の成績を達成するなど、応用においても成果を上げた。

（2）詳細

研究テーマ A「プラットフォームの構築」

プラットフォームの構築については、国際標準に準拠しつつデータ型階層の定義をはじめとするバックエンドの機能を設計することが一つの課題であった。また、グラフィカルユーザインターフェース（GUI）を適切にデザインすることで、自動化・省力化という大きな目標をより洗練されたレベルで達成した。

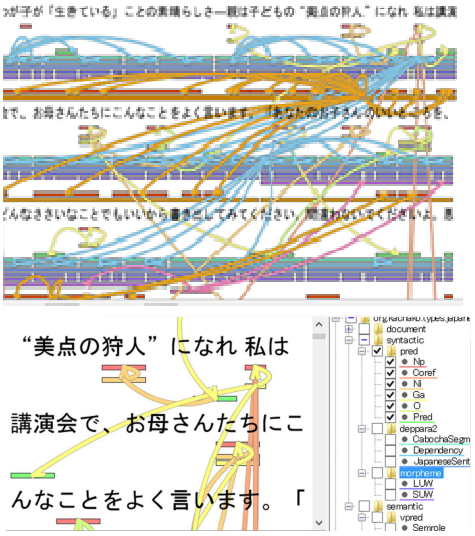

図. 汎用視覚化機能の表示例
BCCWJ の各種アノテーションを同時表示（上図）通常はフィルタや検索などで絞り込んで表示する(下図)

テキストとアノテーションの汎用大規模同時検索について以下で詳述する。自然言語処理においては、他の分野より複雑で、相互に依存関係のあるデータ構造（アノテーション）を扱う傾向にある。本研究で構築した Kachako プラットフォームにおいても同様であり、得られた出力に対して検索機能を実現することは非常に重要であった。そこで、国立国語研究所との共同研究として、Kachako プラットフォームに統合された、アノテーション構造とテキストを同時にクエリ指定できる検索システムを開発した。実装は Apache Solr をベースとして安定性とスケーラビリティを実現し、クエリは領域代数を拡張したものを適用可能とすることで汎用性を担保した。

研究テーマ B「ツールキットの構築と応用」

ツールキットの構築は、下記で各個に詳述するように当初計画よりも様々な研究協力に発展した。以下で説明するツールはいずれも、統合プラットフォーム Kachako に互換であり、統合されている。また、いずれも再利用性を高めるために、できるだけ小さなコンポーネントに分割して実装、再構成している。

国立情報学研究所を中心とする「ロボットは東大に入れるか（東ロボ）」プロジェクトでは、東京大学に合格しうる人工知能の開発をめざし、まずは大学入試センター試験の自動解答に挑戦している。この東ロボプロジェクトに参画し、解答器の開発基盤およびセンター試験社会科の自動解答器構築を行った。社会科解答器については代ゼミ模試タスクにおいて最高の成績を達成した。

国立国語研究所では、大規模均衡日本語コーパス（BCCWJ）を開発している。BCCWJ は

基本となるテキストデータに加え、複数の共同研究チームにより各種の言語学的アノテーションが各個に独立して付与されている。具体的には、形態素・品詞・係り受け・述語項構造などである。国語研究所との共同研究を通じて、これらのテキスト・アノテーションを互換化・標準化したうえで重ね合わせ、Kachako プラットフォームに読み込むコンポーネントを開発した。

質問応答システムは、Apple 社の Siri に代表されるように近年需要の高いアプリケーションであるが、日本語で利用可能なものはすくなく、質問応答システムを Kachako プラットフォームに追加することは重要な研究要素であった。国立情報学研究所との共同研究として、横浜国大およびカーネギーメロン大学の協力を得て、複数の質問応答システムのコンポーネント化と互換化、視覚化ツールの実装を行った。これら Factoid 型の日本語質問応答システムを用いて国際コンテスト型ワークショップ NTCIR-11 QALab タスクを開催した。

本領域 2 期生の駒谷研究者と、音声言語処理に関する研究協力を行い、音声認識システムおよびその言語モデル生成を Kachako 対応コンポーネントとして再構築した。

研究テーマ C「自然な言語処理モデルの研究」

より自然な言語処理を実現しうるモデルの研究のため、国立情報学研究所・国立障碍者リハビリテーションセンター等と協力して視線追跡装置を用いた心理言語学的な実験を行った。これにより、文読解時の視線停留パターンと文章の性質、被験者の認知能力などとの相関がわかり、言語モデル構築のための知見を得ることができた。

3. 今後の展開

本研究は自然言語処理の資源・ツールを「使う」ことの全自動化を目指し、国際標準準拠の自動化支援プラットフォームの構築と、プラットフォーム上で利用可能な言語資源・ツールの収集・構築・互換化を行ってきた。今後は、「作る」「チューニング」「カスタマイズ」のサポートをより充実させるため、プラットフォーム機能への機械学習手法の統合を進めたい。コンポーネントについては、コーパスリーダーと評価メトリクスを中心に、さらに多言語・多ドメインの言語資源・ツールの組み込みを目指したい。また、システムのメンテナンスやサポート、ユーザコミュニティの構築といった研究者のみでは難しい課題を、企業との連携を視野に入れて進めていきたいと考えている。これにより、最終的に社会還元、社会実装が達成できるものと考える。

4. 評価

（1）自己評価

（研究者）

自然言語処理の自動化を支援するプラットフォームの構築と、プラットフォーム上で動作する互換ツールキットの構築は、いずれも当初計画を上回る規模と内容で達成できた。これはひとえにさきがけ研究のサポートあればこそであり、領域内外の共同研究が多く生まれたのもさきがけ研究に採択されたことが大きい。一方で自然な言語モデルの構築は、プラットフォームとツールキットの構築に注力したこと、また研究を進めるにつれテーマそのものがより長期の研究を必要とすることがわかってきたことから、今後の研究の発展が課題である。

研究体制としては、研究者自身の作業が研究の中核であり、さきがけの個人型研究そのもののスタイルで研究を推進してきた。

本研究におけるシステムの構築は、要素要素それぞれが一つの研究テーマをなしうる規模のものであると同時に、それらを統一的、調和的にひとつのシステムにまとめるという野心的な計画であったが、現実にそのようなシステム構築に成功した。すなわち、データやツールの組合せによる自動ワークフロー生成、自動インストール、自動実行、メトリクスのカスタマイズ可能な比較評価統計、汎用視覚化、テキストとアノテーションの同時検索といった機能を、国際標準 UIMA に準拠しつつ統一的な実行プラットフォーム"Kachako"として構築した。一貫して自動化という究極的なユーザ視点から、個々の機能とその統合に必要な設計とシステム構築の研究を行い、個別の機能においても統合システム全体としても、世界的にも他に類を見ない独自のシステムを実現できた。

またツールキットの構築においては、質問応答・音声認識・大学入試自動解答など多様なツールを互換ツールキットとして統合した。プラットフォーム上でのツールキットの実装とその応用により、構築したシステムの実用性を示すことができた。

本研究の成果は、実社会での需要に直結している。産業、研究、教育のいずれの分野においても大きな需要が見込めるため、そのような需要に対応できる運用の仕組みが構築できれば社会への波及効果は非常に大きい。また、本研究の成果をコアとする共同研究もさまざまな分野で拡大しており、いずれも直接的な実社会での応用が見込まれるテーマとなっている。今後さらに研究を発展させ、社会実装につなげていきたい。

（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

（研究総括）

本研究では、研究者が持つ卓越した実装力を遺憾なく発揮して、自然言語処理に関わるワークフローの自動生成、自動インストール、自動実行、評価メトリクスの視覚化といった機能を統合し、自然言語処理プラットフォーム「Kachako」を実現した。さらに、他のさきがけ研究者や、国立情報学研究所を中心とする「ロボットは東大に入れるか（東ロボ）」プロジェクトなど、多くの共同研究を通じて、プラットフォームの有効性を実証的に示してきた。その成果が認められ、国際会議の招待講演や国際会議での受賞などが増えてきている。今後は自然言語処技術を活用する研究のハブとなることを通じて、世界に影響を与える活躍を期待したい。自然言語処理プラットフォームと互換ツールキットの構築について、統合システムの実装を達成している。共同研究を通じた技術の発展や、今後の社会・経済への波及効果が期待できる成果である。

5. 主な研究成果リスト

（1）論文（原著論文）発表

著者．発表論文タイトル．掲載誌名．発行年，巻号，始頁-終頁，その他

1. Yoshinobu Kano. Kachako: a Hybrid-Cloud Unstructured Information Platform for Full Automation of Service Composition, Scalable Deployment and Evaluation. In the 1st

	International Workshop on Analytics Services on the Cloud (ASC), the 10th International Conference on Services Oriented Computing (ICSOC 2012). pp.72-84. Shanghai, China, November 12th 2012.
2.	Yoshinobu Kano. Towards automation in using multi-modal language resources: compatibility and interoperability for multi-modal features in Kachako. In the 8th International Conference on Language Resources and Evaluation (LREC 2012), pp. 1098-1101. Istanbul, Turkey, May 23rd, 2012
3.	Yoshinobu Kano. Kachako: Towards a Data-Centric Platform for Full Automation of Service Selection, Composition, Scalable Deployment and Evaluation. In the International Conference of Web Services (ICWS 2012), pp. 642-643. Hawaii, USA, June 24th, 2012.
4.	Tadayoshi Hara, Chen Chen, Yoshinobu Kano, Akiko Aizawa. Modeling Comma Placement in Chinese Text for Better Readability using Linguistic Features and Gaze Information. In the The Second Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2013), ACL 2013 Workshop. pp. 49-58. The National Palace of Culture, Sofia, Bulgaria. August 8th, 2013.
5.	狩野 芳伸. 統合研究基盤:質問応答システムの互換コンポーネント化による再利用性向上と開発自動化支援. 人工知能学会誌,27(5) 特集号「ロボットは東大に入れるか」, pp.492-495. 2012 年 9 月

(2)特許出願

研究期間累積件数:0 件

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

[国際会議招待講演] Yoshinobu Kano. Kachako for Fully Automated NLP: from workflow creation to large scale parallel processing with ready-to-use NLP toolkit. In Conference of the Pacific Association for Computational Linguistics (PACLING 2013). An invited talk. September 3rd, 2013. Keio University, Tokyo.

[受賞] Best Paper Award, International Workshop on Analytics Services on the Cloud, International Conference on Service Oriented Computing. (筆頭著者として) 2012

[講演発表] 狩野芳伸.「暗記」と人間の知的な処理(社会科). ロボットは東大に入れるか 2013 ー東ロボくん、代ゼミ模試に挑戦!ー 代々木ゼミナール本部校 代ゼミタワー. 2013 年 11 月 23 日

[講演発表] 狩野芳伸, 増田勝也. テキストとアノテーションの汎用同時検索システム. 第 6 回コーパス日本語学ワークショップ. 国立国語研究所. 2014 年 9 月 9 日.

[メディア] ロボットが東大目指す 人に近い人工知能とは. 生放送ゲスト出演. 深層 NEWS, BS 日テレ. 日テレタワースタジオ. 2014 年 2 月 26 日